Hadoop-Benchmark: Rapid Prototyping and Evaluation of Self-Adaptive Behaviors in Hadoop Clusters (Artifact)*

Bo Zhang[1], Filip Křikava[2], Romain Rouvoy[3], and Lionel Seinturier[4]

1 **University of Lille / Inria, Villeneuve d'ascq, France**
bo.zhang@inria.fr

2 **Northeastern University, Boston, MA, USA**
f.krikava@neu.edu

3 **University of Lille / Inria, Villeneuve d'ascq, France**
romain.rouvoy@inria.fr

4 **University of Lille / Inria, Villeneuve d'ascq, France**
lionel.seinturier@inria.fr

Abstract

Arising with the popularity of Hadoop, optimizing Hadoop executions has grabbed lots of attention from research community. Many research contributions are proposed to elevate Hadoop performance, particularly in the domain of self-adaptive software systems. However, due to the complexity of Hadoop operation and the difficulty to reproduce experiments, the efforts of these Hadoop-related research are hard to be evaluated.

To address this limitation, we propose a research acceleration platform for rapid prototyping and evaluation of self-adaptive behavior in Hadoop clusters. It provides an automated manner to quickly and easily provision reproducible Hadoop environments and execute acknowledged benchmarks. This platform is based on the state-of-the-art container technology that supports both distributed configurations as well as standalone single-host setups. We demonstrate the approach on a complete implementation of a concrete Hadoop self-adaptive case study.

1998 ACM Subject Classification Elicitation methods

Keywords and phrases Hadoop, Docker, Rapid Prototyping, Benchmark

Digital Object Identifier 10.4230/DARTS.3.1.1

Related Article Bo Zhang, Filip Křikava, Romain Rouvoy and Lionel Seinturier, "Hadoop-Benchmark: Rapid Prototyping and Evaluation of Self-Adaptive Behaviors in Hadoop Clusters", in Proceedings of the 12th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS 2017).
http://dx.doi.org/10.1109/SEAMS.2017.15

Related Conference 12th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS 2017), May 22–23, 2017, Buenos Aires, Argentina

* This work is partially supported by the Datalyse project: www.datalyse.fr. Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

1 Scope

Hadoop-Benchmark is an open-source research acceleration platform for rapid prototyping and evaluation of self-adaptive behaviors in Hadoop clusters. The main objectives are to allow researchers to

- rapidly prototype, *i.e.*, to experiment with self-adaptation in Hadoop clusters without the need to cope with low-level system infrastructure details,
- reproduction, *i.e.*, to share complete experiments for others to reproduce them independently,
- repetition, *i.e.*, to experiment with and to compare their work, re-doing the same experiments on the same system using the same evaluation methods.

It uses docker and docker-machine to easily create a multi-node cluster (on a single laptop or in multiple-hosts) and provision Hadoop. It contains a number of acknowledged benchmarks and one scenario consisting of some self-adaptive behaviors [1].

The cluster provisioning and benchmark execution is done in an automated way based on simple configuration files which can be easily shared. Furthermore, the provisioned nodes in a cluster include monitoring service that can be used for developing touchpoints for system identification and the monitoring part of feedback control loops governing the self-adaptation.

It is important to note that while Hadoop has been mostly connected with implementation of MapReduce paradigm, there has been an overhaul on its architecture and, since version 2, Hadoop has become a general framework for distributed large-scale applications. Our focus on Hadoop goes therefore beyond MapReduce and has wide applications to other technologies that are based the core enabling technologies—*i.e.*, distributed files-systems (HDFS) and application scheduler (YARN).

2 Content

The Hadoop Benchmark Platform is essentially a set of Docker images and scripts that orchestrate the provisioning and the execution of Docker hosts and containers deployed on these hosts. There are three types of images:

- `scenarios/vanilla-hadoop/` : *Base Images* that provide a vanilla Hadoop installation,
- `scenarios/self-balancing-example/` : *Extension Images* to the base images with custom configuration coupled with implementation of some self-adaptive behavior,
- `benchmarks` : *benchmark images* that execute particular benchmark suites.

and

- `cluster.sh` : one bash script which provides set of options and commands to simplify the creation of Hadoop cluster into only several clicks.

3 Getting the artifact

The artifact endorsed by the Artifact Evaluation Committee is available free of charge on the Dagstuhl Research Online Publication Server (DROPS). In addition, the artifact is also available at: `https://github.com/Spirals-Team/hadoop-benchmark/`.

4 Tested platforms

Hadoop-Benchmark has been successfully tested on Linux (Ubuntu 14.04) and OSX. Hadoop-Benchmark is a docker-based tool. Therefore, the *docker-engine* is necessary:

- *docker*, *version* >= 1.12
- *docker-machine*, *version* >= 0.8

Hadoop-Benchmark has provided several simple R scripts to facilitate the analysis of results generated by acknowledged benchmarks. The R language and several R packages are optionally required:

- *R*, *version* >= 3.3.2
- *R packages*: *tidyverse* and *Hmisc*

In standalone mode, the provisioned Hadoop cluster is installed in multiple VMs based on VirtualBox (*version* >= 5.1).

In multiple-nodes mode, the Hadoop cluster can be deployed to multiple cloud providers [1] including Grid5000[2].

The further information can be found in the tutorial of Hadoop-Benchmark: `https://github.com/Spirals-Team/hadoop-benchmark/wiki/Tutorial`.

5 License

The artifact is available under the Apache License, Version 2.0 (the "License").

6 MD5 sum of the artifact

52f4c2aff782ee91f18b8a2f6fbf845c

7 Size of the artifact

176 KB

References

1 Bo Zhang, Filip Křikava, Romain Rouvoy, and Lionel Seinturier. Self-Balancing Job Parallelism and Throughput in Hadoop. In *16th IFIP International Conference on Distributed Applications and Interoperable Systems (DAIS)*, volume 9687 of *Proceedings of DAIS'16*, pages 129–143, Heraklion, Crete, Greece, June 2016. Springer. `doi:10.1007/978-3-319-39577-7\_11`.

[1] `https://docs.docker.com/machine/drivers/`

[2] For running on Grid5000, Hadoop-Benchmark requires docker-machine-driver-g5k (`https://github.com/Spirals-Team/docker-machine-driver-g5k`)